Out[1]:

Out[2]: Click here to toggle on/off the Python code.

PHYS 10792: Introduction to Data Science

2019-2020 Academic Year

Course instructors: Rene Breton (http://www.renebreton.org) - Twitter @BretonRene (https://twitter.com/BretonRene)
Marco Gersabeck (http://www.hep.manchester.ac.uk/u/gersabec) - Twitter @MarcoGersabeck (https://twitter.com/MarcoGersabeck)

Chapter 11

Syllabus

1. Probabilities and interpretations
2. Probability distributions
3. Parameter estimation
4. Maximum likelihood + extended maximum likelihood
5. Least square, chi2, correlations
6. Monte Carlo basics
7. Probability
8. Hypothesis testing
9. Confidence level
10. Goodness of fit tests
11. **Limit setting**
12. Introduction to multivariate analysis techniques

Topics

[11 Limit setting](#)

[11.1 Coverage](#)

[11.2 The issue of flip-flopping](#)

[11.3 The Feldman-Cousins method](#)

[11.4 The CLs method](#)

11 Limit setting

We discussed confidence leves in Chapters 8 and 9. These were mostly discussed in scenarios not affected by limitations of the parameters that are being assessed. In this part we cover the somewhat more complicated situation when physical boundaries apply, e.g. that a quantity has to be positive (e.g. a mass).

This Chapter largely (11.1-11.3) follows a paper by Feldman and Cousins: Phys. Rev. D57 (1998) 3873 (https://journals.aps.org/prd/abstract/10.1103/PhysRevD.57.3873). This is an excellent paper well worth a read also at your level.

Below is a hypothetical confidence belt based on Gaussian probability density functions with varying mean and width. In this example the parameters are

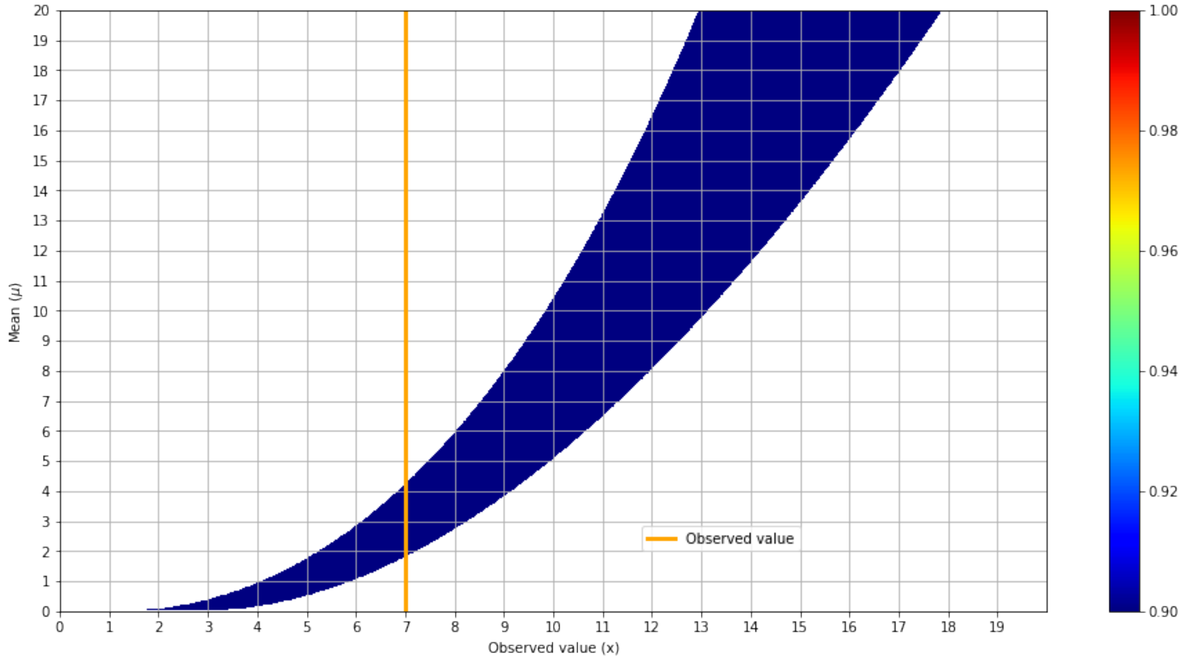
mean $= 2 + 3 \times \sqrt{\mu}$,

and

width $= 0.5 + 0.05 \times \mu$.

As we are now dealing with a continuous variable, the confidence level covered by each horizontal band is exactly 90%.

A problem arises if we want to assign a confidence interval for μ for a measured value around $x = 1$.



11.1 Coverage

The construction of confidence belts, which we discussed previously, is based on defining horizontal intervals according to a certain confidence level C. These can be constructed as central confidence intervals according to

$$P(x < x_1|\mu) = P(x > x_2|\mu) = (1 - C)/2,$$

or as upper confidence limit intervals

$$P(x < x_1|\mu) = 1 - C.$$

For a given measured value of x_0, these then lead to an interval for μ with

$$P(\mu \in [\mu_1, \mu_2]) = C.$$

This statement means that the unknown true value of μ, μ_t lies within the interval $[\mu_1, \mu_2]$ in a fraction C of the experiments conducted.

This has to be distinguished from the statement that the degree of belief that μ_t lies in $[\mu_1, \mu_2]$ is C. This is a Bayesian statement for $P(\mu_t|x_0)$ for which we need to know the prior for μ_t according to Bayes' theorem

$$P(\mu_t|x_0) = P(x_0|\mu_t)P(\mu_t)/P(x_0).$$

Returning to the previous equation, this being satisfied indicates that the intervals *cover* μ at the stated confidence, or that they have the correct *coverage*.

Alternatively, the case when

$$P(\mu \in [\mu_1, \mu_2]) < C,$$

is called *undercoverage*. This is a serious issue as this implies the existence of more information than is the case in reality.

Conversely,

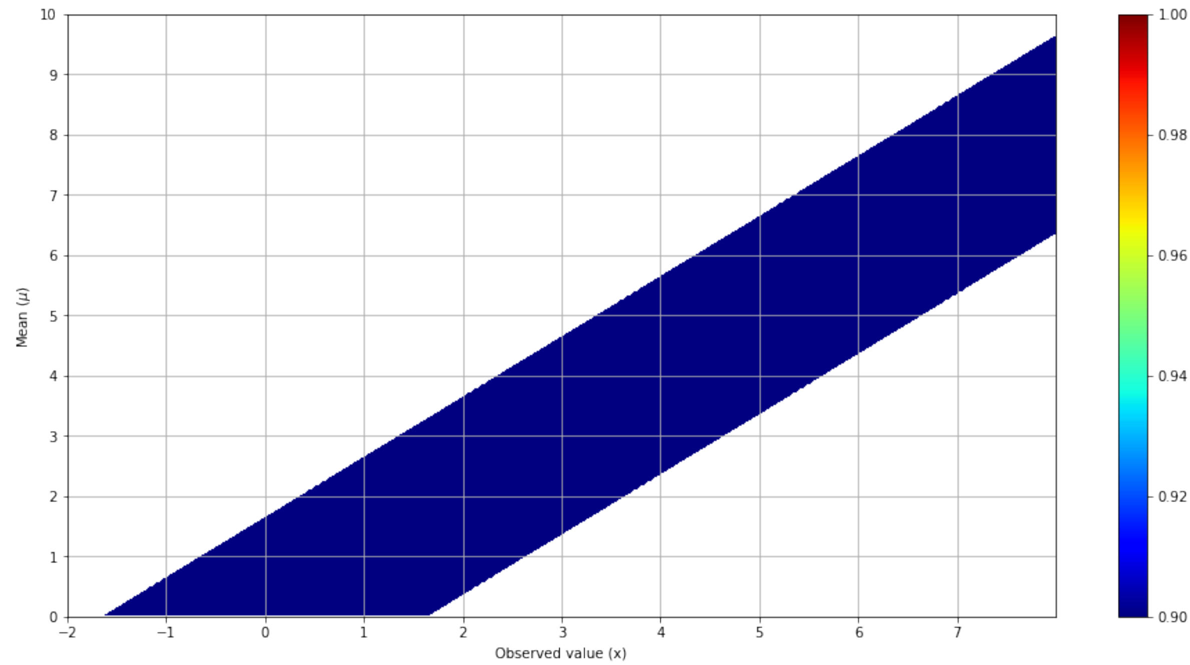
$$P(\mu \in [\mu_1, \mu_2]) > C,$$

is called *overcoverage* , which is in most cases a less severe issue, but it leads to a loss of rejection power against false hypotheses. This case is sometimes called conservatism. Being conservative is never optimal and one should always strive to make estimates as accurate as possible.

Recap: Confidence levels from Gaussians

In the case of Gaussian distribution functions, the construction becomes very simple. The x_- and x_+ curves become straight lines and the limits are obtained simply by $\mu_\pm = x_0 \pm n\sigma$, where $n = 1$ for 68% confidence level, $n = 1.64$ for 90% confidence level, and so on.

This is a simplified case of the introductory example above.



Recap: Measurement of a constrained quantity

The measurement of a mass was mentioned in the introduction to confidence levels.

Assuming this measurement occurs with Gaussian uncertainties, we have the probability distribution function

$$P(x|\mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}.$$

The construction with a confidence belt fails miserably when a negative mass is measured with a relatively small uncertainty, e.g. (-0.5 ± 0.2) g. When trying to construct a 90% confidence interval, we would get $\mu_\pm = (-0.5 \pm 0.2 \times 1.64)$ g, i.e. even $\mu_+ = -0.172$ g remains negative.

A solution to this issue is a Bayesian construction with a normalisation that takes the physical limit into account. For example for positive masses, one gets

$$P(\mu|x) = \frac{e^{-(x-\mu)^2/2\sigma^2}}{\int_0^\infty e^{-(x-\mu')^2/2\sigma^2} d\mu'} (x < 0).$$

This construction will then lead to one limit being zero, i.e. we set an upper limit.

We will now discuss the limitations of this approach.

Limitations of central confidence belts

Let us continue to consider the example of a Gaussian probability density function.

If we decide *a priori* to construct central intervals, we get the confidence belt already shown a couple of slides ago.

This belt achieves correct coverage for all allowed values of μ. However, as the measured value x_0 gets more and more negative the corresponding interval $\mu_\pm$ tends towards the empty set, which is reached for $x_0 \leq -1.64\sigma$.

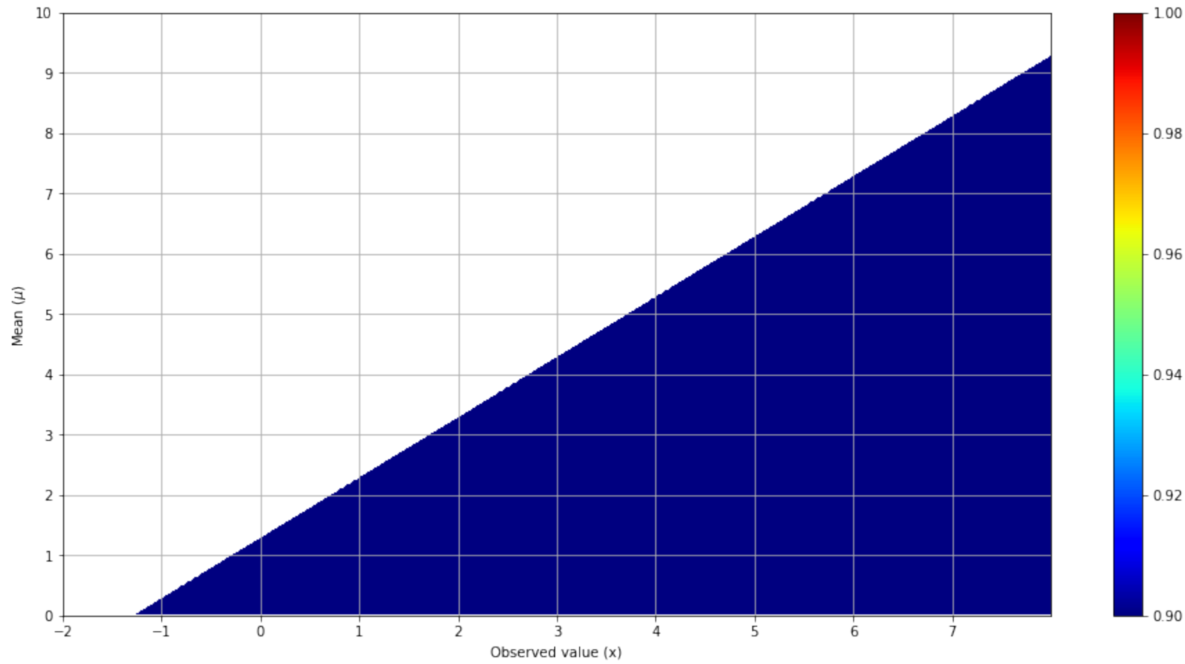
This is clearly an unsatisfactory result. One alternative might be to continue the confidence belt to negative values of μ, but this is similarly unsatisfactory as obtaining a set $\mu_\pm$ with a negative μ_- does not really make sense as we would know that any negative values of μ are unphysical.

Furthermore, the probability distribution function $P(x|\mu)$ may be ill-defined for negative μ.

Limitations of upper confidence intervals

We can similarly to the previous case set out *a priori* to construct upper limit intervals.

In this case the same problem arises for negative measured values with the continuation to negative values of μ being equally unsatisfactory.



11.2 The issue of flip-flopping

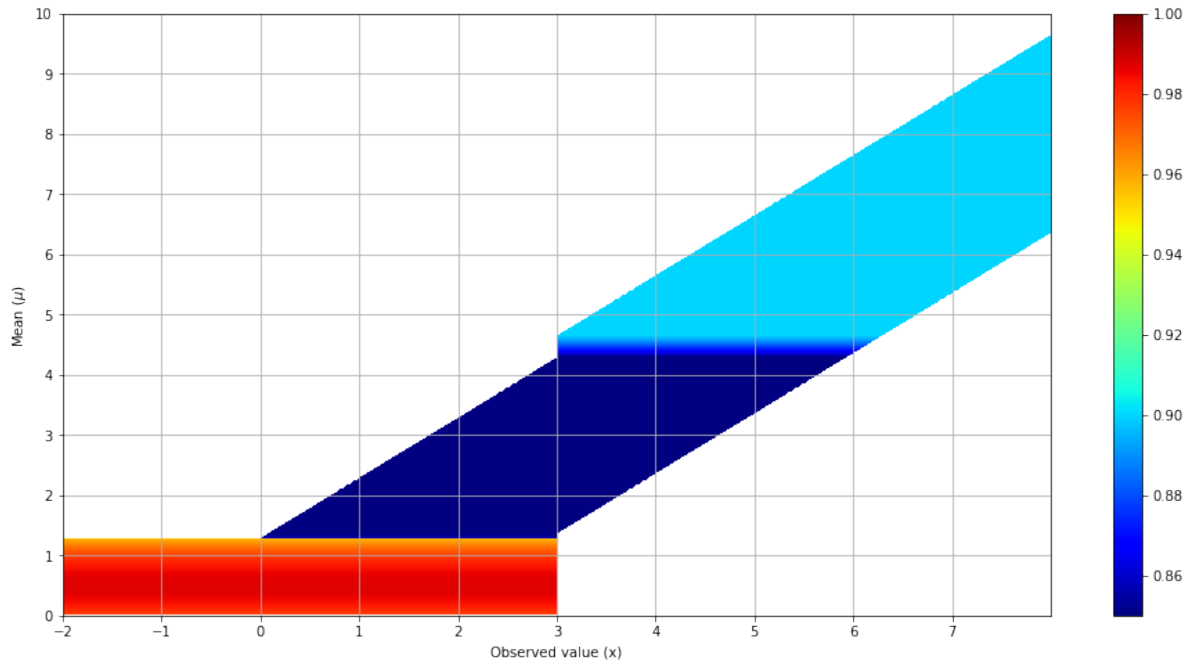
Combining different approaches

As a remedy one might decide to combine these approaches, for example according to the following recipe:

- For a measured value above 3σ we quote a central interval.
- For a less significant value we quote an upper limit.
- For a negative measured value we quote a constant upper limit, the same as for $x_0 = 0$.

Do you think that this is a useful and statistically sound approach?

menti.com (https://menti.com) with key 94 83 82



Flip-flopping

You should have received sufficient warnings about fully defining your procedure prior to performing the measurement that this measurement-dependent recipe should set off some alarm bells.

Let's have a look at the coverage of this approach.

For $\mu = 2$ the coverage is too low as the right-hand limit is defined by the central interval belt, which appears to be shifted up compared to the upper limit belt. Therefore, the intervals undercover, which is problematic.

Also, considering $\mu = 1$, it is apparent that the intervals overcover. More generally, for μ between 0 and about 1.2 the coverage obviously varies as the interval stays constant while $P(x|\mu)$ changes.

Poisson belts

In the case of a counting experiment we have to use Poisson intervals, which have the particularity of leading to stair-like plots as they refer to a discrete distribution, i.e. we can only observe whole events.

In such experiments we are often interested in identifying a particular signal in the presence of background events.

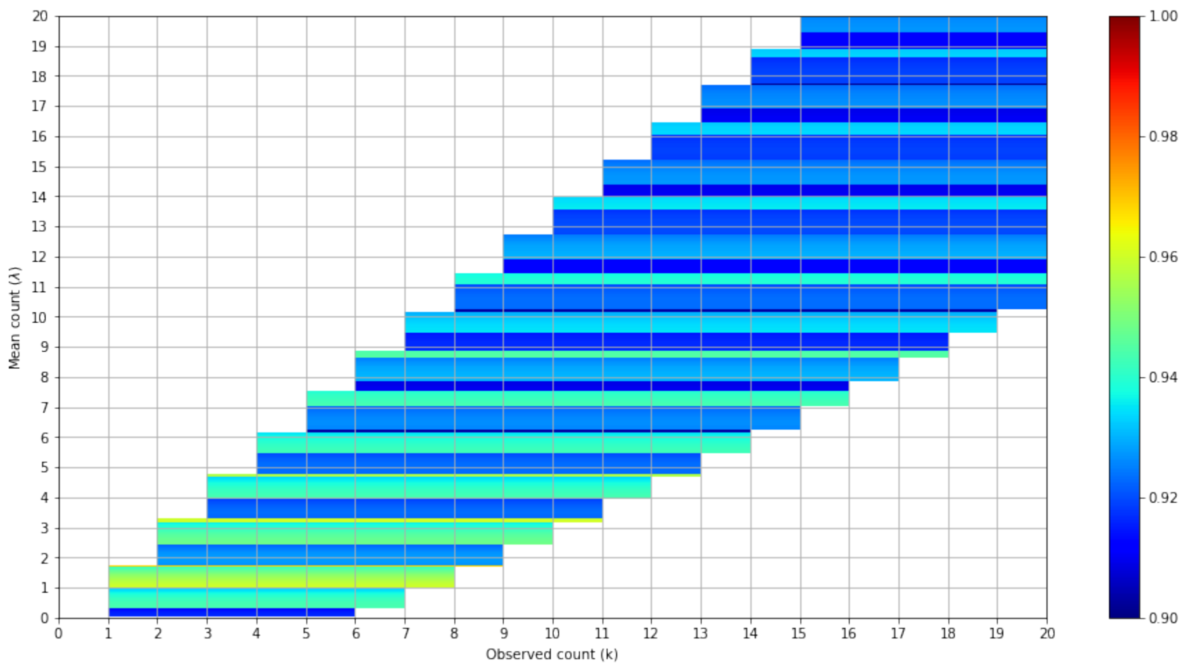
The Poisson distribution function in the presence of a known amount of background, b, is

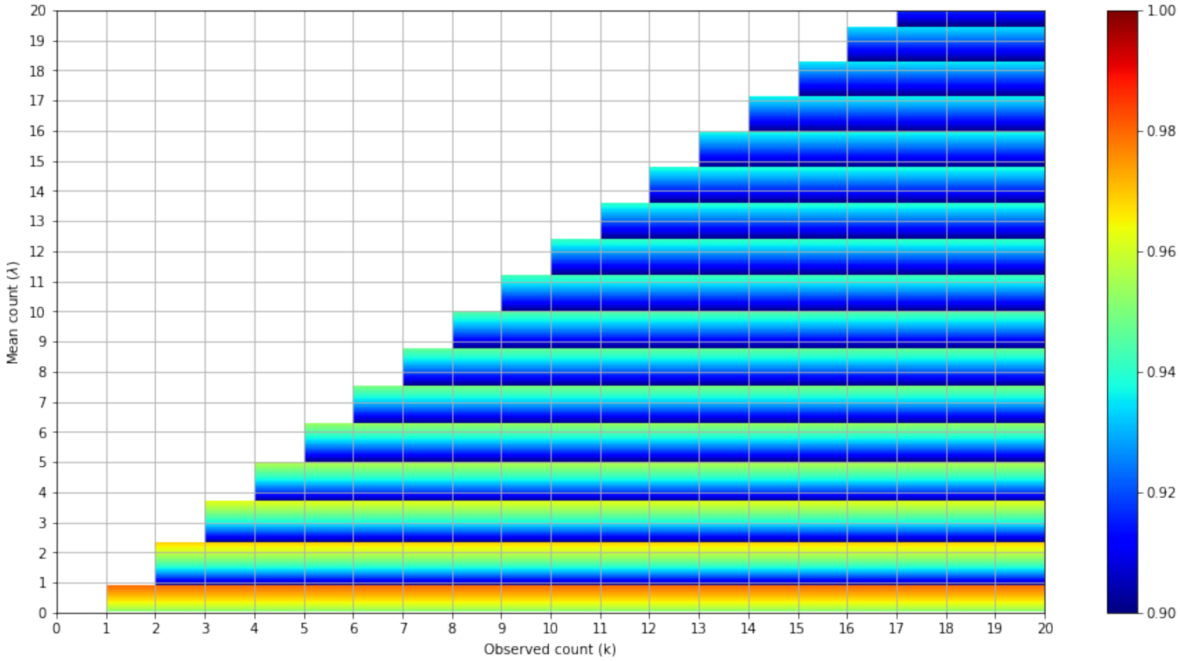
$$P(n|\mu) = e^{-(\mu+b)}\frac{(\mu+b)^n}{n!}.$$

As discussed before, when defining confidence intervals we rather overcover as we are in general unable to hit the desired confidence level exactly.

The following plots show the resulting confidence belts for central intervals and upper limits. They assume $b = 3$.

The same issues apply as for the Gaussian case.




Mean count (λ)
Observed count (k)
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
1.00
0.98
0.96
0.94
0.92
0.90

11.3 The Feldman-Cousins method

The issue of flip-flopping is caused by the mixture of vertical and horizontal construction of the confidence intervals. It is therefore desirable to have an approach that solely depends on a horizontal construction.

A purely horizontal construction cannot a priori define whether it results in an interval or an upper limit as this is, in the end, a result of a vertical cut through the confidence belt at the measured value.

Feldman and Cousins proposed the following alternative approach. The method may appear a little contrived, but this is simply due to the fact that it is strictly restricted to a horizontal construction that does not pre-define whether a central interval or an upper limit is being constructed.

The following example is based on a counting experiment (therefore involving Poisson distributions) with the presence of background. This means that the total count n will include a certain number b of background events, where b is assumed to be known.

It proceeds as follows:

- Consider the construction of a confidence interval of level α for a given μ.
- For each value of n there is a value of μ that maximises $P(n|\mu)$ and this value shall be called μ_{best}.
- In our case this is simply $\mu_{\text{best}} = \max(0, n - b)$.
- Define the ratio $R = P(n|\mu)/P(n|\mu_{\text{best}})$. Given that μ_{best} maximises the probability distribution function, this ratio is between 0 and 1.
- Order all possible values of n by their corresponding values of R and, starting from the largest R, add the values of n to the confidence interval until the sum of their probabilities exceeds α.

The relevant values for an example with $b = 3$ and $\mu = 0.5$ are given in the table below. The probabilities of rank 1-7 add up to (compare the column 'CumulativeSum')

$$0.189 + 0.216 + 0.185 + 0.132 + 0.106 + 0.030 + 0.077 = 0.935,$$

which covers a 90% interval.

This construction leads to an automatic transition between one-sided and two-sided intervals and thereby gives a pre-defined transition from quoting a two-sided interval to an upper limit. Note that nowhere in this procedure does the observed number of events enter.

There are two caveats around the Feldman-Cousins method:

- In the case of having observed 0 events, a measurement with greater levels of expected background can lead to a lower upper limit compared to a measurement with fewer expected background events.
- The approach does not readily allow the inclusion of so-called nuisance parameters, which are used to account for systematic uncertainties. However, alternatives exist that mitigate this issue.

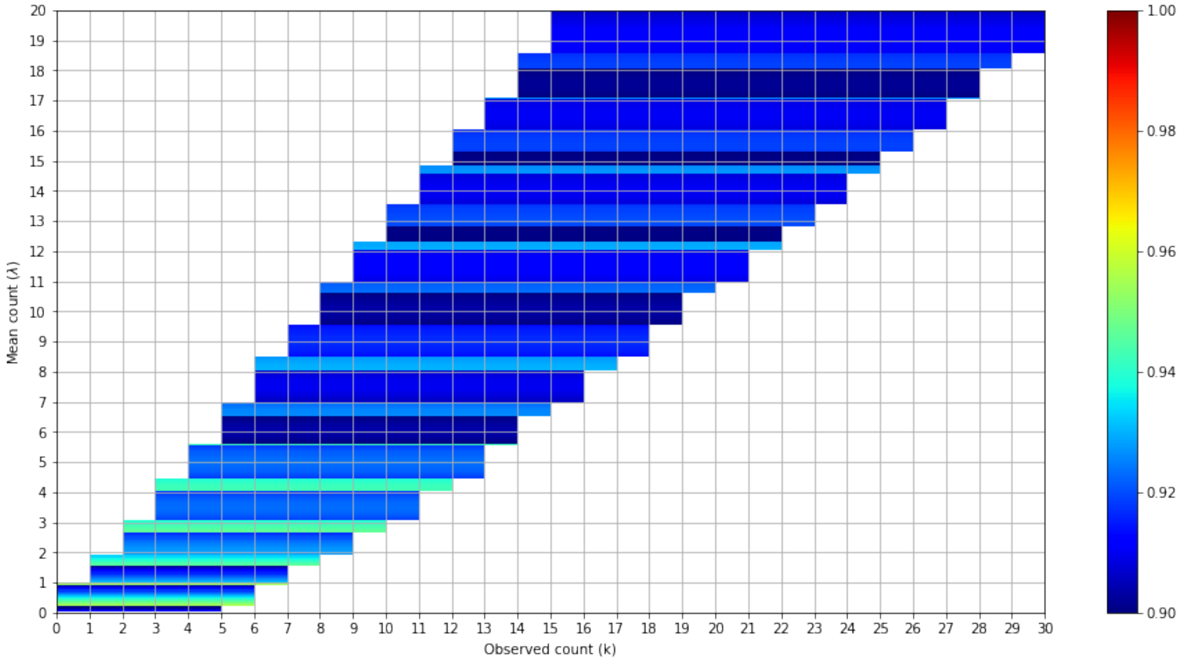
The corresponding confidence belt looks as follows.

```
  1%|          | 2/399 [00:00<00:30, 12.85it/s]

Calculating P(n|mu_best)
Calculating intervals

  3%|▎         | 12/399 [00:00<00:27, 14.06it/s]
```

Table for mu = 0.5

	k	P(k|mu)	mu_best	P(k|mu_best)	R	CumulativeSum
4	4.0	1.888123e-01	1.0	0.195367	9.664501e-01	0.188812
3	3.0	2.157855e-01	0.0	0.224042	9.631482e-01	0.404598
2	2.0	1.849590e-01	0.0	0.224042	8.255556e-01	0.589557
5	5.0	1.321686e-01	2.0	0.175467	7.532375e-01	0.721725
1	1.0	1.056908e-01	0.0	0.149361	7.076191e-01	0.827416
0	0.0	3.019738e-02	0.0	0.049787	6.065307e-01	0.857614
6	6.0	7.709835e-02	3.0	0.160623	4.799953e-01	0.934712
7	7.0	3.854917e-02	4.0	0.149003	2.587145e-01	0.973261
8	8.0	1.686526e-02	5.0	0.139587	1.208230e-01	0.990126
9	9.0	6.558714e-03	6.0	0.131756	4.977938e-02	0.996685
10	10.0	2.295550e-03	7.0	0.125110	1.834825e-02	0.998981
11	11.0	7.304022e-04	8.0	0.119378	6.118396e-03	0.999711
12	12.0	2.130340e-04	9.0	0.114368	1.862708e-03	0.999924
13	13.0	5.735530e-05	10.0	0.109940	5.216973e-04	0.999981
14	14.0	1.433883e-05	11.0	0.105989	1.352858e-04	0.999996
15	15.0	3.345726e-06	12.0	0.102436	3.266167e-05	0.999999
16	16.0	7.318776e-07	13.0	0.099218	7.376494e-06	1.000000
17	17.0	1.506807e-07	14.0	0.096285	1.564950e-06	1.000000
18	18.0	2.929902e-08	15.0	0.093597	3.130327e-07	1.000000
19	19.0	5.397188e-09	16.0	0.091123	5.922961e-08	1.000000
20	20.0	9.445079e-10	17.0	0.088835	1.063212e-08	1.000000
21	21.0	1.574180e-10	18.0	0.086712	1.815420e-09	1.000000
22	22.0	2.504377e-11	19.0	0.084733	2.955602e-10	1.000000
23	23.0	3.811008e-12	20.0	0.082884	4.597981e-11	1.000000
24	24.0	5.557721e-13	21.0	0.081152	6.848574e-12	1.000000
25	25.0	7.780809e-14	22.0	0.079523	9.784356e-13	1.000000
26	26.0	1.047417e-14	23.0	0.077989	1.343037e-13	1.000000
27	27.0	1.357762e-15	24.0	0.076540	1.773928e-14	1.000000
28	28.0	1.697203e-16	25.0	0.075169	2.257850e-15	1.000000
29	29.0	2.048348e-17	26.0	0.073869	2.772941e-16	1.000000
30	30.0	2.389740e-18	27.0	0.072635	3.290088e-17	1.000000

```
100%|██████████| 399/399 [00:24<00:00, 16.10it/s]
```



Continuous case

For a Gaussian (i.e. continuous) variable the procedure is very similar. In short, it is:

- For a Gaussian distributed variable μ with boundary condition $\mu \geq 0$, find the value of μ for which $P(x|\mu)$ is maximal and call this μ_{best}. This is given by $\mu_{\text{best}} = \max(0, x)$.
- Define the likelihood ratio

$$R(x) = \frac{P(x|\mu)}{P(x|\mu_{\text{best}})}.$$

- For a given μ find the interval $[x_1, x_2]$ such that $R(x_1) = R(x_2)$ and $\int_{x_1}^{x_2} P(x|\mu)dx = \alpha$, with α the desired confidence level. This integral gives the accepted range for each value of μ.

The condition $R(x_1) = R(x_2)$ replaces the explicit ordering and, provided that $R(x)$ is continuously falling on either side of $x = \mu_{\text{best}}$, it uniquely defines the interval $[x_1, x_2]$ for a given α.

At $x = 1.28$ the distribution transitions from a one-sided to a two-sided interval.

The full confidence belt is shown in the following plot.

Computation requirements

The continuous case requires calculating and sorting many values in x for each value in μ. The exact number is defined by the desired resolution of the plot.

The plot below already takes several minutes to compute while it is based on a simple Gaussian function. However, here we knew the correct values for μ_{best} a priori.

If μ_{best} needs to be calculated, this can add considerable computation requirements.

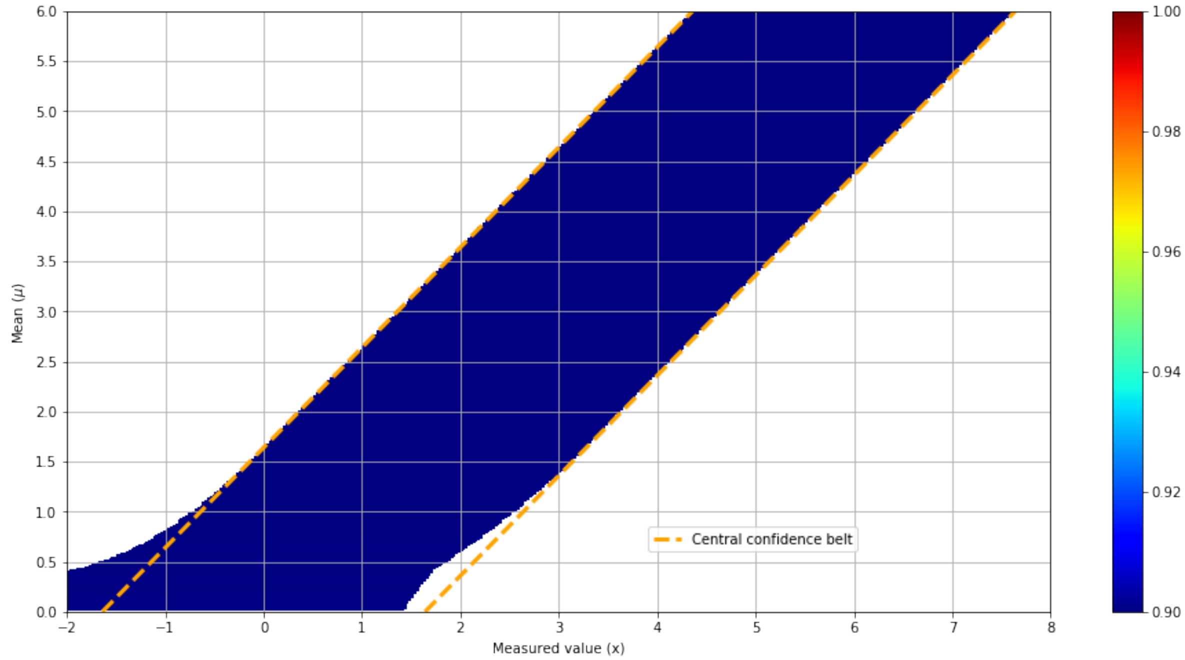
In reality, this method is likely to be applied to much more complex probability densitiy functions, which take significantly longer to evaluate. Therefore, it can be computationally challenging to obtain a result with the Feldman-Cousins method.

```
  0%|          | 0/399 [00:00<?, ?it/s]

Calculating P(n|mu_best)
Calculating intervals

  1%|▏         | 5/399 [00:03<03:55,  1.67it/s]

Table for mu =  0.075
```

	x	P(x|mu)	mu_best	P(x|mu_best)	R	CumulativeSum
83	0.075	3.989423e-01	0.075	0.398942	1.000000e+00	0.009974
82	0.050	3.988176e-01	0.050	0.398942	9.996875e-01	0.019944
84	0.100	3.988176e-01	0.100	0.398942	9.996875e-01	0.029914
81	0.025	3.984439e-01	0.025	0.398942	9.987508e-01	0.039876
85	0.125	3.984439e-01	0.125	0.398942	9.987508e-01	0.049837
80	0.000	3.978218e-01	0.000	0.398942	9.971915e-01	0.059782
86	0.150	3.978218e-01	0.150	0.398942	9.971915e-01	0.069728
79	-0.025	3.969525e-01	0.000	0.398818	9.953235e-01	0.079652
87	0.175	3.969525e-01	0.175	0.398942	9.950125e-01	0.089575
78	-0.050	3.958377e-01	0.000	0.398444	9.934590e-01	0.099471
88	0.200	3.958377e-01	0.200	0.398942	9.922179e-01	0.109367
77	-0.075	3.944793e-01	0.000	0.397822	9.915980e-01	0.119229
76	-0.100	3.928800e-01	0.000	0.396953	9.897405e-01	0.129051
89	0.225	3.944793e-01	0.225	0.398942	9.888130e-01	0.138913
75	-0.125	3.910427e-01	0.000	0.395838	9.878865e-01	0.148689
74	-0.150	3.889708e-01	0.000	0.394479	9.860359e-01	0.158414
90	0.250	3.928800e-01	0.250	0.398942	9.848041e-01	0.168236
73	-0.175	3.866681e-01	0.000	0.392880	9.841888e-01	0.177902
72	-0.200	3.841389e-01	0.000	0.391043	9.823452e-01	0.187506
71	-0.225	3.813878e-01	0.000	0.388971	9.805050e-01	0.197040
91	0.275	3.910427e-01	0.275	0.398942	9.801987e-01	0.206816
70	-0.250	3.784198e-01	0.000	0.386668	9.786683e-01	0.216277
69	-0.275	3.752403e-01	0.000	0.384139	9.768350e-01	0.225658
92	0.300	3.889708e-01	0.300	0.398942	9.750052e-01	0.235382
68	-0.300	3.718551e-01	0.000	0.381388	9.750052e-01	0.244679
67	-0.325	3.682701e-01	0.000	0.378420	9.731788e-01	0.253885
66	-0.350	3.644919e-01	0.000	0.375240	9.713558e-01	0.262998
65	-0.375	3.605270e-01	0.000	0.371855	9.695362e-01	0.272011
93	0.325	3.866681e-01	0.325	0.398942	9.692332e-01	0.281678
64	-0.400	3.563824e-01	0.000	0.368270	9.677200e-01	0.290587
..	...	...	...	...	...	...
370	7.250	2.642618e-12	7.250	0.398942	6.624061e-12	0.981581
371	7.275	2.207990e-12	7.275	0.398942	5.534610e-12	0.981581
372	7.300	1.843692e-12	7.300	0.398942	4.621450e-12	0.981581
373	7.325	1.538538e-12	7.325	0.398942	3.856543e-12	0.981581
374	7.350	1.283089e-12	7.350	0.398942	3.216226e-12	0.981581
375	7.375	1.069384e-12	7.375	0.398942	2.680548e-12	0.981581
376	7.400	8.907158e-13	7.400	0.398942	2.232693e-12	0.981581
377	7.425	7.414353e-13	7.425	0.398942	1.858503e-12	0.981581
378	7.450	6.167880e-13	7.450	0.398942	1.546058e-12	0.981581
379	7.475	5.127754e-13	7.475	0.398942	1.285337e-12	0.981581
380	7.500	4.260366e-13	7.500	0.398942	1.067915e-12	0.981581
381	7.525	3.537491e-13	7.525	0.398942	8.867175e-13	0.981581
382	7.550	2.935434e-13	7.550	0.398942	7.358041e-13	0.981581

383	7.575	2.434321e-13	7.575	0.398942	6.101937e-13	0.981581
384	7.600	2.017492e-13	7.600	0.398942	5.057103e-13	0.981581
385	7.625	1.670992e-13	7.625	0.398942	4.188557e-13	0.981581
386	7.650	1.383138e-13	7.650	0.398942	3.467014e-13	0.981581
387	7.675	1.144156e-13	7.675	0.398942	2.867975e-13	0.981581
388	7.700	9.458750e-14	7.700	0.398942	2.370957e-13	0.981581
389	7.725	7.814670e-14	7.725	0.398942	1.958847e-13	0.981581
390	7.750	6.452323e-14	7.750	0.398942	1.617358e-13	0.981581
391	7.775	5.324148e-14	7.775	0.398942	1.334566e-13	0.981581
392	7.800	4.390487e-14	7.800	0.398942	1.100532e-13	0.981581
393	7.825	3.618294e-14	7.825	0.398942	9.069719e-14	0.981581
394	7.850	2.980051e-14	7.850	0.398942	7.469879e-14	0.981581
395	7.875	2.452855e-14	7.875	0.398942	6.148396e-14	0.981581
396	7.900	2.017664e-14	7.900	0.398942	5.057533e-14	0.981581
397	7.925	1.658648e-14	7.925	0.398942	4.157614e-14	0.981581
398	7.950	1.362662e-14	7.950	0.398942	3.415688e-14	0.981581
399	7.975	1.118796e-14	7.975	0.398942	2.804405e-14	0.981581

```
[400 rows x 6 columns]

100%|██████████| 399/399 [04:10<00:00,  1.59it/s]
```



11.4 The CLs method

The CL_s method is another limit-setting method commonly used in high-energy physics. It has some coverage issues that will be discussed later, which is why it is disliked by some purists, but it has other practical advantages over stricter methods such as Feldman-Cousins.

The method is based on comparing two hypotheses. The null hypothesis is that an observation is based on background only, while the alternative hypothesis is that an observation is due to a combination of background and signal.

One defines the two probabilities, the first

$$p_{s+b} = P(x \leq x_{\text{obs}}|s+b) = \int_{-\infty}^{x_{\text{obs}}} f(x|s+b)dx,$$

which represents the probability for observing an x that is more background-like than x_{obs}. For a counting experiment this corresponds to $x \leq x_{\text{obs}}$. The value of p_{s+b} depends on the amount of signal assumed in the hypothesis.

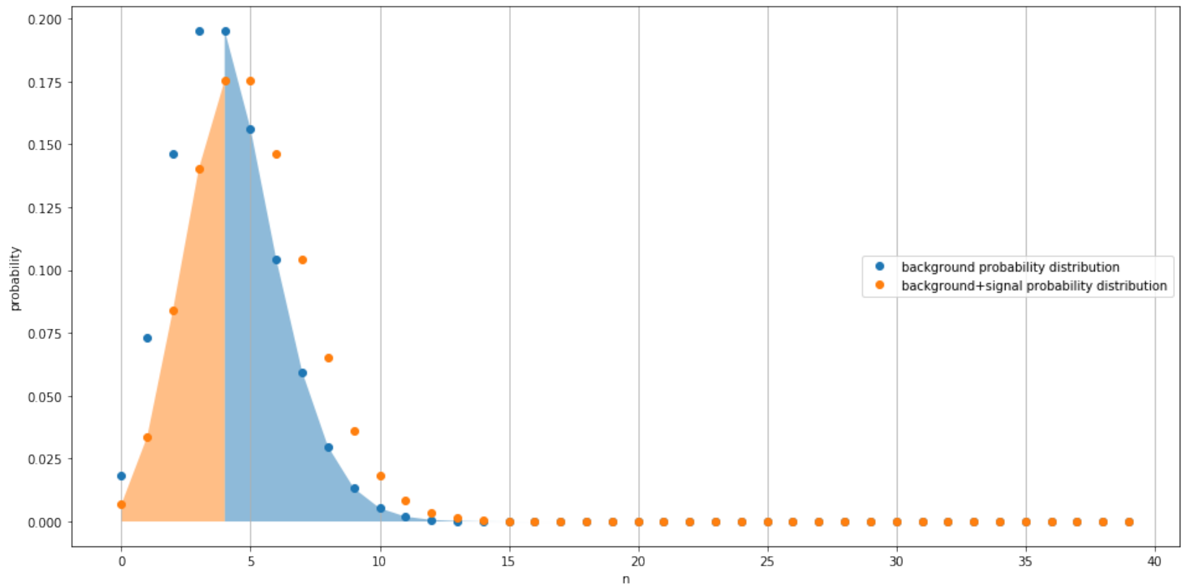
The second probability is

$$p_b = P(x \geq x_{\text{obs}}|b) = \int_{x_{\text{obs}}}^{\infty} f(x|b)dx,$$

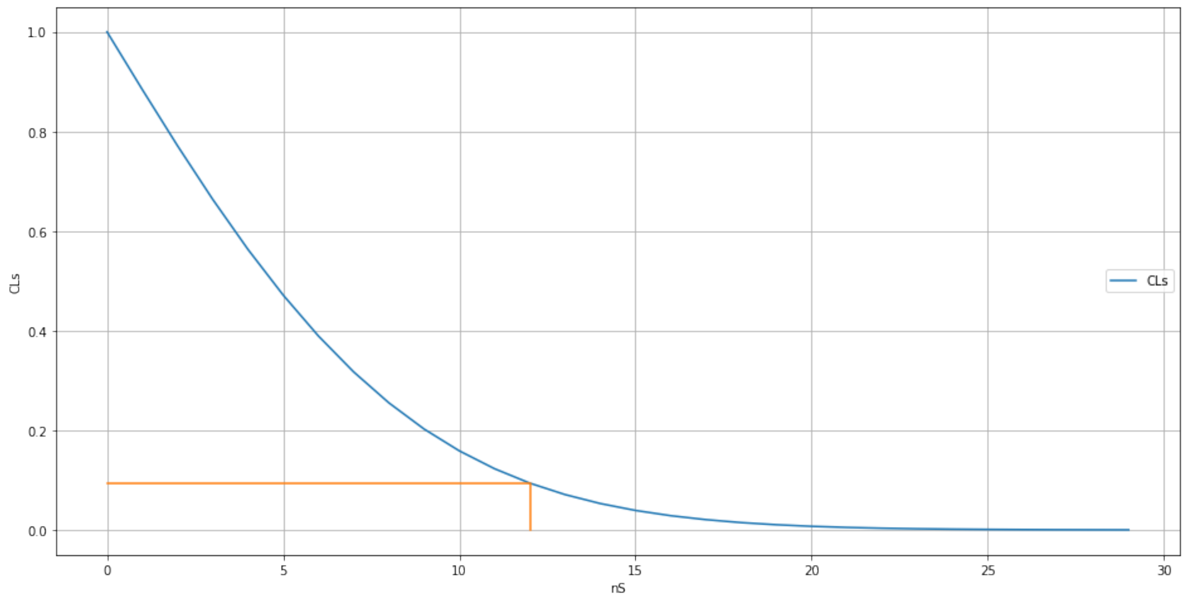
which represents the probability of a pure background sample giving a value of x that is more signal-like than x_{obs}.

The following gives an illustration of the situation.

```
Working with nB = 4 , nS = 1 , and an observed count of 4
p_(s+b)=0.4405
p_b=0.3712
p_(s+b)/(1-p_b)=0.7005
```



```
The 90.0% CL limit is 12
```



Limit setting

We discussed confidence leves in weeks 7 and 8. These were mostly discussed in scenarios not affected by limitations of the parameters that are being assessed. In this part we cover the somewhat more complicated situation

Considering for a moment only the first probability, for a small signal compared to the amount of background (as illustrated here) and a very small observed value, i.e. a downward fluctuation in the background, leads to a limit with low significance. In the example above an observation of 1400 counts leads to a p-value of 0.4%.

However, there is 99.5% of the background-only spectrum that looks more signal-like than the observed value of 1400 counts.

The CL_s method modifies the initial p-value and defines

$$CL_s \equiv \frac{p_{s+b}}{1-p_b}.$$

This fraction effectively introduces a penalty term for the case that both hypotheses have very similar probability distribution functions, which is equivalent with saying that the measurement has little sensitivity. We already discussed exactly this aspect when covering hypothesis testing.

In our example, CL_s now becomes $\frac{0.4\%}{1-99.5\%} = \frac{0.4\%}{0.5\%} = 80\%$, so we would not be able to set any stringent limit due to the lack of sensitivity.

In the extreme case of a measured count of zero, i.e. neither signal nor background lead to a measurement, CL_s is always 100%, which then prevents the setting of a 90% upper limit. While this is considered a mathematical drawback by some, it actually represents a meaningful safety mechanism against setting unjustified limits.

If the two hypotheses have very well separated distributions, a signal-like observation carries practically no penalty term.

Example: how many students in this course have already a university degree?

menti.com (https://menti.com) with key 55 63 86.

CL_s in practice

When applying the CL_s method in practice one would want to establish the sensitivity of a method beforehand. Therefore, one would calculate the CL_s values for a range of conceivable amounts of signal and for the expected level of background.

As the amount of background will fluctuate according to Poisson statistics, what is commonly done is to simulate an ensemble of background sets and calculate the CL_s value for each of them. Then, one plots the median CL_s value for every considered signal value alongside the spread of the ensemble.

Finally, once the number of observed counts have been determined, one can draw a curve of observed limits with the actual signal and background numbers. The ensemble of expected limits now serves as a control check against statistical fluctuations.

The single value to be quoted as the observed limit depends on the desired confidence level.

Here is an example of a recent publication by our group. Reference: R. Aaij et al. (LHCb collaboration), Phys. Lett. B754 (2016) 167 (https://doi.org/10.1016/j.physletb.2016.01.029). The limit is set at the 90% confidence level (hence the line at $CL_s = 0.10$) and the observed limit is below the median expected limit due to a downward fluctuation in the observed background.